Autonomous Robots (2023) 47:1119–1135
https://doi.org/10.1007/s10514-023-10107-7

An overview of space-variant and active vision mechanisms for resource-constrained human inspired robotic vision

Rui Pimentel de Figueiredo[1] · **Alexandre Bernardino**[1]

Received: 22 May 2021 / Accepted: 2 May 2023 / Published online: 9 June 2023

Abstract
In order to explore and understand the surrounding environment in an efficient manner, humans have developed a set of space-variant vision mechanisms that allow them to actively attend different locations in the surrounding environment and compensate for memory, neuronal transmission bandwidth and computational limitations in the brain. Similarly, humanoid robots deployed in everyday environments have limited on-board resources, and are faced with increasingly complex tasks that require interaction with objects arranged in many possible spatial configurations. The main goal of this work is to describe and overview biologically inspired, space-variant human visual mechanism benefits, when combined with state-of-the-art algorithms for different visual tasks (e.g. object detection), ranging from low-level hardwired attention vision (i.e. foveal vision) to high-level visual attention mechanisms. We overview the state-of-the-art in biologically plausible space-variant resource-constrained vision architectures, namely for active recognition and localization tasks.

Keywords Biologically inspired vision · Active vision · Space-variant vision · Selective and divided attention · Object detection

Abbreviations

ACF	Aggregated channel features
AIP	Anterior parietal lobe
ANN	Artificial neural networks
BO	Bayesian optimization
CAD	Computer-aided design
CIP	Caudal Intraparietal Sulcus
CNN	Convolutional neural network
DNN	Deep neural networks
DCNN	Deep convolutional neural network
GMM	Gaussian mixture model
GMM	Gaussian mixture models
GP	Gaussian proccess
GP	Gaussian proccesses
GPU	Graphical proccessing unit
GPU	Graphical proccessing units
GHT	Generalized Hough transform
HOG	Histogram of gradients
HRI	Human robot interaction
HVS	Human visual system
IOR	Inhibition of return
KL	Kullback–Leibler
JPDA	Joint probabilistic data association
LGN	Lateral geniculate nucleus
LSTM	Long short-term memory units
MAB	Multi-armed bandit
MAP	Maximum a posteriori
MCTS	Monte Carlo tree search
MOT	Multiple object tracking
NBV	Next-best-view
POMDP	Partially observable Markov decision process
POMDP	Partially observable Markov decision processes
GMM	Gaussian mixture models
RANSAC	Random sample consensus
RF	Receptive field
RF	Receptive fields
RNN	Recurrent neural network
RPN	Region proposal network
SES	Sensory ego-sphere
SIFT	Scale invariant feature transform

✉ Rui Pimentel de Figueiredo
ruihortafigueiredo@gmail.com

Alexandre Bernardino
alex@isr.tecnico.ulisboa.pt

[1] Instituto Superior Tecnico, Universidade de Lisboa, Lisbon, Portugal

SVM	Support vector machine
SVM	Support vector machines
UCB	Upper confidence bound
UCB	Upper confidence bounds
UT	Unscented transform
WTA	Winner take all
RoI	Region of interest
FPS	Frames per second
MOTP	Multiple object tracking precision

1 Introduction

Space-variant vision and attention mechanisms are the fundamental processes in biological systems, responsible for prioritizing the elements of the visual scene to be attended, i.e., to control perceptual resources (Amso & Scerif, 2015; Parasuraman & Yantis, 1998) and cope with the brain computational limitations. Humans rely on space-variant sensing (foveal vision), and on stimulus-driven (bottom-up) and goal-driven (top-down) information processing mechanisms to define where in the visual input the attentional foci should be oriented to Katsuki and Constantinidis (2014). This way, information processing is constrained and directed towards salient or task-relevant stimuli. Likewise, an important issue in many computer vision applications requiring real-time performance, resides in the involved computational effort (Borji & Itti, 2013b), especially in robotics where energy efficient, fast and accurate perception is a fundamental requirement, e.g., in visual localization and servoing during grasping, manipulation and hand-over of tools to human or machine collaborators. In humanoid robotics, in particular, real-time operation is conditioned by physical limitations on on-board computational and power resources, as well as data transmission bandwidth if one opts to outsource information processing to outside servers. Therefore much effort has been made towards understanding the underlying principles of biological attention mechanisms and applying those mechanisms in robotics, in an attempt to build more efficient solutions, capable of performing in real-time, under resource-constrained settings (Begum & Karray, 2011).

We overview works on space-variant low-level vision (i.e. foveal vision) to higher level perception, i.e., selective attention mechanisms.

The main topics over-viewed in this work, can be summarized as follows:

(1) Neural and artificial mechanisms of visual information processing;
(2) Computational models for foveal vision mechanisms
(3) Computational models of selective visual attention
(4) Biologically plausible methods for active object localization and recognition;

Table 1 Outline of the different models for biologically inspired vision over-viewed in this article. From low-level physiological mechanisms, to higher level cognitive ones, finalizing with applications in robotics contexts

	Sections
Foveal vision	
Geometric	2.2.1
Filtering	2.2.3
Pyramid	2.2.2
Learning	2.2.4
Attention and Spatial Selectivity	
Bottom-up	2.4
Top-down	2.4
Hybrid	2.4
Visual Tasks	
Classification	3
Detection	3
Applications	
Humanoid Robotics	4

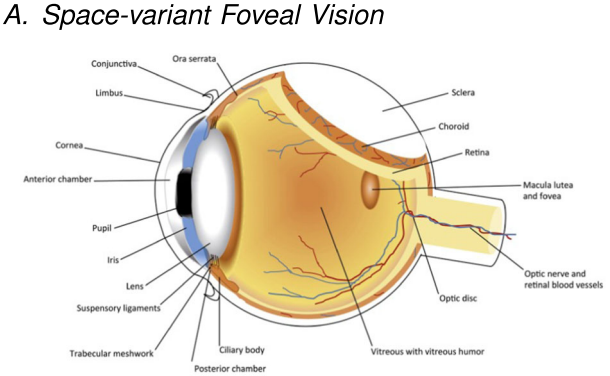
(5) Applications of the former mechanisms and computational models in humanoid robotics.

In the remainder of this article we overview the neurophysiology of the Human Visual System (HVS), and review the state-of-the-art in biologically plausible space-variant vision models, focusing on artificial foveal vision and visual attention mechanisms. This review focuses on highlighting the state-of-the-art methods rather than providing quantitative and qualitative comparisons between methodologies.

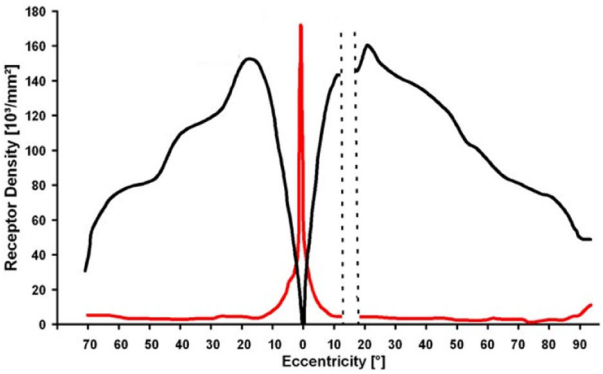
Our review on space-variant vision and attention mechanisms differs from other works (Posch, 2012; Kartheek Medathati et al., 2016; Fernández-Caballero & Ferrández, 2017) by describing in detail the human visual system and linking with classical and modern computational models for artificial foveal vision, selective visual attention, and active vision, focused on object recognition and localization, as well as on implementations in robotics visual setups (see Table 1).

2 Neural and artificial mechanisms of visual information processing

The process of seeing starts with light entering the eye through the cornea. The eye has the ability to adapt to different levels of brightness (adaptation) and to shape its lens and pupil size in order to focus objects at different distances (accommodation). The light passing through the pupil, is focused by the lens, onto the retina, a sensory membrane responsible for receiving and converting the visual stimuli



(a) Eye morphology (figure taken from [9])



(b) Photo-receptors density in the retina. Rods (black) are mostly concentrated in the periphery of the retina and are responsible for low light level vision (scotopic vision). Cones (red) are concentrated in the center of the retina (fovea) and are responsible for high acuity color vision.

Fig. 1 Human eye physiology

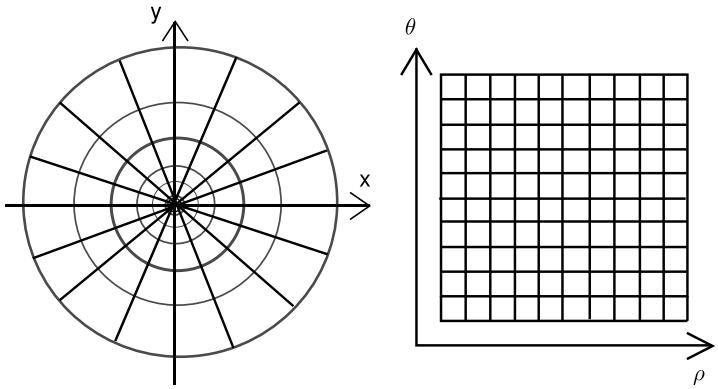
into electric signals to be transmitted to the visual cortex in the brain through the optic nerve (Mohlin et al., 2017).

The retina is mainly composed of two types of photo-receptors: rods which are mostly concentrated at the periphery and are sensitive to brightness and colorless low-light vision (scotopic vision) and the cones that are concentrated mostly in the center of the eye, in a place called fovea, and are responsible for high acuity color vision (see Fig. 1). Finally, the visual signals entering through the optic nerve reach the back of the brain, where the visual cortex is located and the stimuli interpreted.

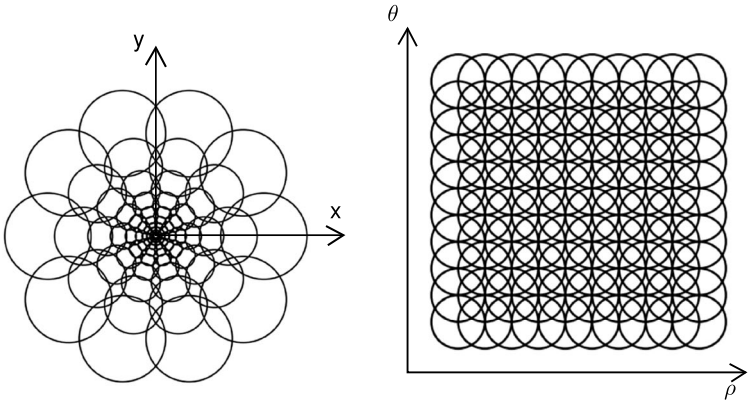
2.1 Space-variant foveal vision

Unlike uniform vision provided by conventional imaging sensors, human vision is space-variant, due to the uneven organization of the photo-receptors in the retina. Visual acuity, provided by the cones, is highest at the fovea, located in the center of the retina, and declines monotonically towards the periphery, with increasing eccentricity (see Fig. 1). This space-variant resolution perception phenomenon—named foveation—is a hardwired mechanism and a natural way of reducing the amount of information streamed to the brain, in order to cope with power, neuronal transmission bandwidth limitations, and the brain machinery processing capacity. In fact, if foveal resolution visual stimuli across the whole field of view was to be processed, the human brain weigh would be significantly increased [to approximately 60 kg (Balasuriya & Siebert, 2005)]. However this compression phenomenon introduces a space-variant uncertainty in visual processes. In order to efficiently explore and understand the surrounding environment (Posner, 2012), humans have developed a set of attention and oculo-motor mechanisms, namely saccades, that allow them to actively and sequentially direct their eyes towards different regions of interest in the surrounding environment, and thus, to cleverly compensate for the aforementioned limitations.

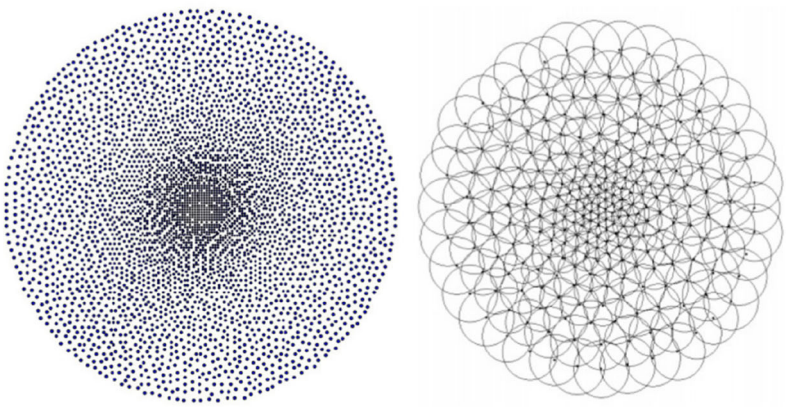
Similar to humans, robots deployed in everyday environments, are faced with increasingly complex scenarios where objects are arranged in many possible different spatial configurations. The problem of deciding which regions in the visual field are to be attended during visual search tasks is computationally demanding or even intractable if approximate solutions are not considered (Tsotsos, 1990). Therefore, like biological systems, humanoid robots must be endowed with mechanisms to allow them to locate objects of interest and to sequentially build detailed representations of the scene, while avoiding the potential overload of processing irrelevant sensory stimuli. Under the assumption that biological systems perform quasi-optimally in their environment due to multiple generations of genetic improvement, researchers have been developing robotics systems (Metta et al., 2008) provided with biologically inspired space-variant image processing (Schwartz et al., 1995; Javier Traver & Bernardino, 2010), gaze control models (Roncone et al., 2016; Bernardino & Santos-Victor, 1999) and attention systems (Begum & Karray, 2011; Borji & Itti, 2013b; Vijayakumar et al., 2001; Frintrop et al., 2010; Potapova et al., 2017). These implementations not only mimic the mechanisms observed in humans but, in general, also lead to more efficient and effective behaviors under resource-constrained settings (bandwidth, computational and energetic). In the context of robotics, and from a practical standpoint, unconventional space-variant sensing representations, in particular human-like foveal vision, offer multiple advantages when compared to conventional uniform counterparts, including reduced resolution with wide field-of-view, being suitable for real-time performance in active vision systems (Bajcsy et al., 2018; Schwartz et al., 1995) that are able to manipulate the view-point and other visual parameters.



(a) Retinal (left) and cortical (right) log polar representations with non-overlapping superpixel Receptive Fields (RFs).



(b) Retinal (left) and cortical (right) log-polar representations with overlapping circular RFs. Left: the x and y correspond to Cartesian coordinates in the retinal plane. While ρ and θ correspond to coordinates in the cortical domain.



(c) Self-organized Gaussian receptive field tessellation produced with self-similar neural network units. Left: node tessellation. Right: Gaussian receptive fields on top of a retina tessellation.

Fig. 2 Log-polar transform

2.2 Computational foveal vision mechanisms

All levels of the visual system are highly regular and symmetric, from the photoreceptors distribution in the retina, to higher-level cell organization in the striate cortex. Different digital sensing architectures exist in the literature that attempt to mimic biological vision structures, namely adaptive and reconfigurable hardware-based ones (García et al., 2014; Bailey & Bouganis, 2009), as well as algorithmic-based human like vision ones (Almeida et al., 2018).

Biologically plausible foveated digital image processing techniques attempt to mimic the space-variant phenomena in the visual pathways, and have numerous applications, including video streaming in low-bandwidth networks (e.g. teleoperation and remote surveillance) and scene understanding tasks (e.g. object detection (Akbas & Eckstein, 2017),

tracking (Bernardino & Santos-Victor, 1999; Gould et al., 2007), and robot navigation (Santos-Victor & Bernardino, 2003)). The algorithms proposed in the literature, try to mimic foveal vision and can be classified as geometric (Javier Traver & Bernardino, 2010), multi-resolution (Adelson et al., 1984), filtering-based (Geisler & Perry, 1998; Wang, 2003), or learning-based (Lukanov et al., 2021; Cheung et al., 2017).

2.2.1 Geometric-based approaches

Studies from neurophysiology have shown that the receptive field spacing and size scale exponentially with eccentricity in the retina, and that light stimuli produces activation displacements in the cortex that are inversely proportional to the distance to the fovea.

Geometric-based approaches attempt to model the retinotopic mapping transformation, using geometric shapes, that occurs between RFs in the retina and the Lateral Geniculate Nucleus (LGN) (Hubel & Wiesel, 1968), where neighboring retinal locations are mapped to neighboring cortical locations. This RFs mapping distribution can be mathematically approximated using the log-polar transformation (Schwartz, 1977), which is given by the following mathematical expression:

$$(\rho, \theta) = \left(log\left(\sqrt{((x - x_c)^2 - (y - y_c)^2)}\right)\right), \; atan\left(\frac{(y - y_c)}{(x - x_c)}\right) \quad (1)$$

and has attracted much interest within the robotics community (see Fig. 2a, b). First, because it allows trading-off field-of-view, resolution and data compression. Second, they provide some degree of invariance to rotations and scaling transformations, as these become linear shifts in the cortical plane.

Many log-polar models have been proposed in the literature (Bolduc & Levine, 1998) and may be categorized as conformal non-overlapping or overlapping, depending on the RF support radius (see Fig. 2). Although being computationally more intensive than their non-overlapping RFs counterparts, overlapping models are better at approximating the space-variant averaging phenonema in the retina, and produce smoother retinal mappings. Still, the literature falls short on works that attempt to model uncertainty in 3D reconstruction due to space-variant quantization phenomena in the retina, and to leverage these uncertainty measures for Next-Best-View (NBV) planning during exploration and visual search tasks (de Figueiredo et al., 2018).

While the previous approaches attempt to capture the retina receptive field tessellation structure through analytic geometric modeling, other approaches capture its underlying structure through exploration and learning strategies. One example is the self-organized retina of Balasuriya (2006) that unlike previous approaches can deal with sampling discontinuities between the fovea and the peripheral region of the visual field. During the structure creation process, they use self-similar neural network units, whose weights undergo random transformations to produce randomized tessellations (see Fig. 2c).

2.2.2 Multi-resolution pyramids

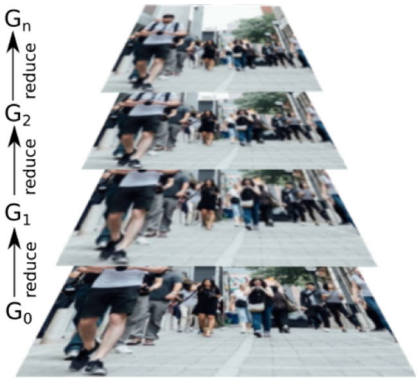
Image pyramids (Adelson et al., 1984) have been proposed for multi-resolution image processing and are particularly suited for multi-scale image analysis, data compression, and as an intermediate step of key point extraction algorithms (e.g. Scale Invariant Feature Transform (SIFT)). The basic principle resides on low-pass smoothing and downsampling.

Gaussian pyramids are the most common in the literature and utilize Gaussian kernels for the smoothing operation. The first level in the pyramid (level 0) contains the original image g_0 that is first low-pass filtered via convolution with $2D$ isotropic and separable Gaussian filter kernels, and then downsampled by a factor of two, yielding the image g_1 at level 1. Successive images g_{k+1} are obtained from the previous levels g_k, by iteratively repeating the low-pass filtering and down-sampling procedures (see Fig. 3a). Gaussian pyramids are useful for many applications, in particular for recognizing patterns of unknown scale (e.g. scale invariant template matching), and for fast foveated coarse-to-fine pattern localization (see Fig. 3b).

The Laplacian pyramid (see Fig. 3c) was first introduced in Burt and Adelson (1983), for image compression, and is constructed by computing differences of Gaussians. During the construction of the pyramid, each level of the Gaussian pyramid g_k is subtracted from an expanded version of g_{k+1}, to ensure similar resolution and obtain a band-pass image L_k. Data compression is achieved by storing the largely decorrelated L_k and the low-pass filtered image g_{k+1}, instead of g_k.

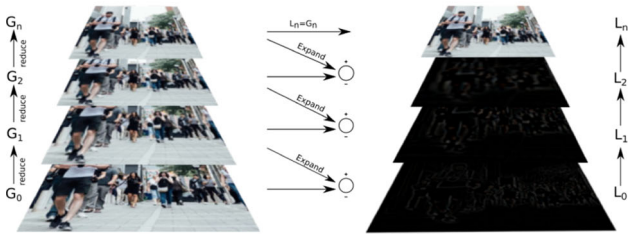
2.2.3 Filtering-based methods

In the work of Geisler and Perry (1998) the authors proposed a foveation mechanism for digital image streaming in low-bandwidth communication channels, that allows the user to point the high spatial resolution focus to regions of interest, with pointing devices (e.g. eye tracker or mouse), being also suitable for studies involving eye movements. The method starts by building a Laplacian pyramid, then, each level is multiplied by an exponential kernel, centered at the foveation point, upsampled and summed with the previous levels, to obtain an image that matches the psychophysical space-variant contrast sensitivity of the HVS (see Fig. 4). Matching the falloff resolution of the HVS, makes optimal



(a) Gaussian pyramid

(b) Foveated low-pass image pyramid



(c) Laplacian pyramid

Fig. 3 Multi-resolution pyramid representations

use of compression resources, by discarding only the details that cannot be resolved by the human eye, via manipulation of the exponential kernel standard deviation. Inspired by this model the authors of Almeida et al. (2018), Melício et al. (2018) developed a real-time implementation that was used to study the impact of artificial foveal vision mechanisms in gaze sequence modelling.

2.2.4 Learning-based methods

More recent learning-based approaches employ deep neural networks, that learn to deploy attention at specific image locations, depending on the task. The approach of (Recasens et al., 2018) proposes a saliency-based distortion layer for convolutional neural networks that is optimized to perform task-dependent spatial sampling of input visual data. The proposed layer learns to deform high-resolution image data by downsampling in a non-uniform and non-linear manner such that task-relevant information is preserved while irrelevant discarded.

Spatial transformer networks (Jaderberg et al., 2015) introduced the ability to learn space-invariant representations, from simple invariance to translations, rotations and scaling to more complex warpings. The similarly minded method of Thavamani et al. (2021), learns to magnify regions in the field-of-view that are likely to enclose objects. These salient regions are processed at high resolution, to ensure high detection accuracy, while keeping computational complexity tractable. The use of differentiable image warping, using spatial transformer networks, ensures bounding box estimations can be mapped back to the original image space. In the work of Lukanov et al. (2021), the authors propose a method in which the input image is foveated with Foveal Cartesian Geometry (FCG) and classified by a CNN. An attention map is computed from the last convolutional layer, that is used for attention using salient features. A Global Average Pooling (GAP) layer is used before the classification output layer, to assist the attention mechanism in augmenting the attention map such that features specific to particular classes of objects are inhibited or prioritized. Finally, the maximum intensity in the attention map is used as the location to which the fovea should move next. The PatchDrop method of Uzkent and Ermon (2020) proposes a reinforcement learning approach that dynamically identifies when and where to acquire high resolution data conditioned on low resolution images.

2.3 Visual attention and spatial selectivity as resource constrained perception

Visual attention is the process through which organisms select a sub-part of the visual stimuli to be processed in detail, while suppressing the rest, to obtain an efficient perception and cope with limited brain computational resources.

The first studies on visual attention date back to the mid 19th century, pioneered by Von Helmholtz (1866) and motivated by the willingness to understand how humans attend stimuli at the periphery of the visual field. By designing a device called tachistoscope Helmholtz demonstrated independence between the ocular attention focus (i.e. gaze location) and the peripheral attentional foci.

The first model for visual attention was proposed by Broadbent (1958), Quinlan and Dyson (2008), in his filter theory, which introduced the structural bottleneck concept (a limitation on the amount of information that the brain can process), that suggests that selective filters are necessary to decide which stimuli to process and which to ignore. Nowa-

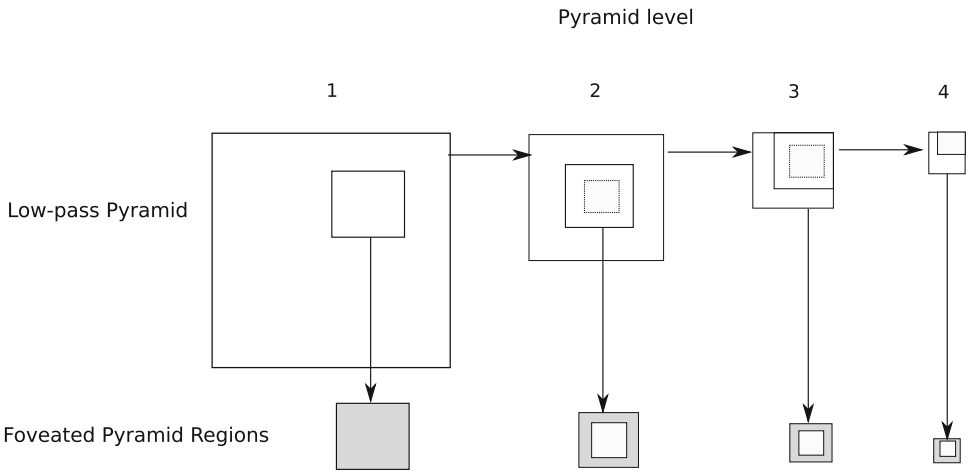


Fig. 4 Filtering-based foveation [see method of Geisler and Perry (1998)]

days, the literature on visual attention is vast, and covers a wide range of scientific fields, including cognitive neuroscience (Carrasco, 2011) and computer science (Borji & Itti, 2013b), playing an important role in computer vision and robotics applications (Begum & Karray, 2011). Attention modeling is not just a multidisciplinary but also a challenging topic under active research due to its importance in controlling the regions (where) and the features or objects (what) the observer should attend to, over time (when). Attention mechanisms can be either selective or divided.

Seminal studies from Hubel and Wiesel (1959, 1962) suggest that the RFs in the mammalian visual cortex increase in size along the visual stream, covering wider areas of the visual field. In parallel, information is selectively processed and the abstraction level of the representations selected along the visual pathways, increase in complexity and in a hierarchical tree manner. Selective attention mechanisms deploy resources to single features or locations, in a serial manner, while divided mechanisms prioritize resources to multiple features or locations, in a parallel manner.

2.3.1 Selective attention mechanisms

Selective visual attention mechanisms are the processes through which biological organisms select only part of the visual signal to be processed while suppressing and ignoring the rest to obtain an efficient perception, and cope with limited neural resources in the brain, allocated to vision. It covers all factors that influence information selection mechanisms, whether they are driven by visual stimuli (bottom-up) or by task-related expectations (top-down) (Bisley, 2011). In particular, spatial attention has been often compared to a spotlight that selectively discards information outside a subarea of the field-of-view. The more sophisticated zoom lens model of Eriksen and St James (1986) suggests that the size of the attentional spotlight is dynamic and object magnification inversely proportional to the lens power (i.e. the spotlight size). Other selective attention theories attempt to explain feature integration (Treisman, 1980), based on Treisman (1985) the idea of determining which visual features are detected preattentively and how the visual system makes the preattentive processing (Treisman, 1980). To identify the preattentive features, (Treisman, 1980) made experiments to detect targets and measuring performance response time and accuracy. In the response time model, viewers were asked to complete the task as quickly as possible and the number of distractors on the display varied. To understand how preattentive processing is done, Treisman proposed a model (see Fig. 5). where each feature map registers the activity of a specific visual feature channel like contrast or size. When an image is shown, features are encoded in parallel into their respective maps. These maps only provide us the activity log of each feature. If the target has a unique feature, we just have to check if there is activity on the respective feature map. However, for conjunction target, one feature map is not enough. Thereby, a serial search must be done in order to find the target that has the correct combination of features. In this case, a focus of attention is used to increase the time and effort spent.

Mishkin et al. (1983) proposed that the visual pathways can be functionally distinguished between *ventral* and *dorsal*, both originating in the primary visual area (V1) (see Fig. 6). The ventral stream mediates feature extraction and object recognition (what) whereas the dorsal stream is specialized in motion and location selectivity (where).

a) Recognition Pathway

Visual stimuli entering the ventral pathway is foveal and neurons within the ventral stream respond selectively to visual features that are important for recognition tasks. Input is grouped in increasingly complex and meaningful visual elements along the pathway. Stimuli selectivity ranges from low-level orientation and color contrast selectivity in V1 and V2, to aggregated contour features and complex shapes in V4 ending in higher-level object representations in the infe-

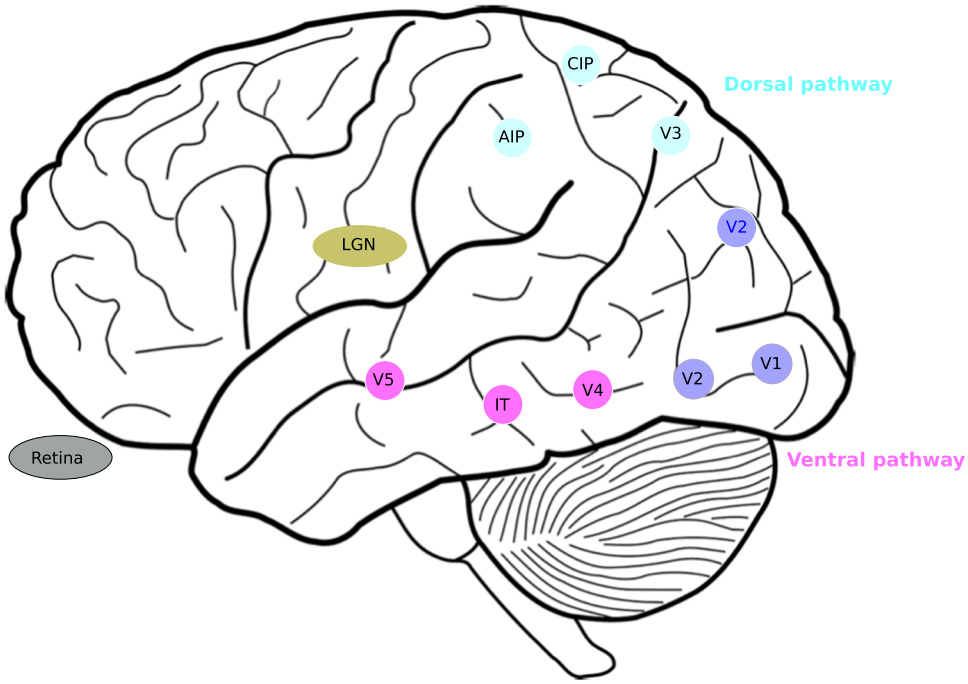


Fig. 6 Human Visual System Pathways: Mishkin et al. (1983) suggested that the visual pathways of primates are organized in two functionally distinct cortical areas (ventral and dorsal), both originating in the primary visual area (V1). The visual stimuli is captured in the retina and is projected into the striate cortex (V1) via the lateral geniculate nucleus of the thalamus (LGN). The ventral stream is responsible for feature extraction and object recognition (what) and the dorsal stream for motion and location selectivity (where)

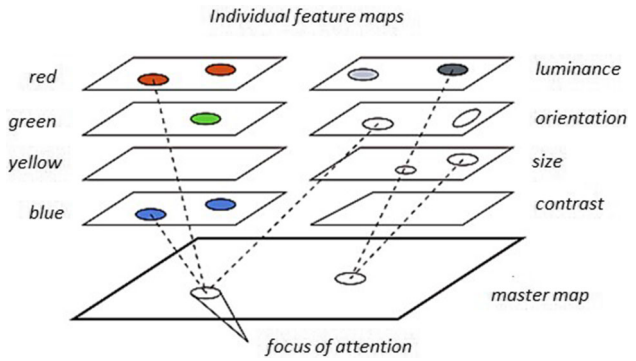


Fig. 5 Treisman's feature integration model of early vision—detection of activity in individual feature maps can be done in parallel, but to search for a combination of features, attention must be focused. Figure adapted from Healey and Enns (2011)

rior temporal (IT) cortex, which comprise category-specific cells. Visual representations are encoded in allocentric or object-centric reference frames. Neurons involved in low-level detection of disparity, were mainly found in the visual cortex, in areas V1, V2 and V3 (Tsutsui et al., 2005), whereas neurons involved in high-level disparity processing facilitate computation of view-point invariant object-specific attributes, to ease recognition functions.

b) Localization Pathway

Neural circuits in the dorsal pathways are tuned for spatial location and motion detection, playing an important role in visuomotor coordination (e.g. in visually guided reaching and grasping). The dorsal stream processes both foveal and peripheral stimulus, and builds a detailed spatial map of object locations and orientations in the field of view. High-level disparity processing, or the reconstruction of 3D surface orientation through the computation of disparity gradients, were found mainly in the Caudal Intraparietal Sulcus (CIP), in the dorsal stream.

In Rosenberg et al. (2013), the authors studied how 3D shape orientation is visually encoded in the brain. In particular, they developed analytical methods to study neural encoding of 3D surface orientation features in the CIP, in the dorsal stream. By varying the orientation of a planar chess pattern positioned frontoparallel with respect to human subjects, the authors concluded that neurons in the CIP jointly encode pan and tilt orientation of 3D surfaces, and that the distribution of preferences over orientations is statistically close to uniform. Nevertheless, it is still unclear if other areas in the brain exhibit unbiased activation selectivity. It is known, however, that areas such as V4 are tuned for specific 3D orientations (Hinkle & Connor, 2002), and that 3D features for grasping and manipulation are context-dependent in the CIP area.

At last, although different neuro computational models have been proposed in the computer vision literature for orientation selectivity in 2D (orientation, motion), it is scarce on works that attempt to model space-variant biases for stimuli selectivity in 3D for enhanced pose estimation (Figueiredo et al., 2019, 2017).

2.4 Computational models of visual attention

James (1980) defined two modes of attention orienting that facilitate the processing and selection of information: stimuli-driven (exogenous) and task-driven (endogenous). The observer attention can be stimuli-driven, triggered by scene characteristics like color or orientation (bottom-up factors) or by specific visual characteristics that depend on the current task or goal (top-down factors). On the one hand, bottom-up processing refers to the involuntary mechanisms responsible for directing resources to salient regions based on differences from a region and its surround (e.g. contrast). In this case, the stimuli directly triggers our attention and, thus, it is a data-driven process. The exogenous system is responsible for orienting our attention, in an involuntary and reflexive manner, to salient locations, features or to where sudden changes occur. For instance, when a light source flashes, ones reaction will be to reflexively direct the gaze to the source (Sokolov & Vinogradova, 1975). On the other hand, top-down processing corresponds to allocating attention voluntarily to features, objects or spatial regions based on prior knowledge and the agent current goals (Posner, 1980). Thus, prior knowledge and the task at hand are used to influence attention in a goal-driven manner. The endogenous mechanisms are voluntary and responsible for directing the attentional resources to predetermined locations, features or objects. Orienting of attention results from taking into account task-specific internal goals, for example, when searching for specific objects or counting how many people will pass through a door. By guiding our attention to task-relevant places we make the counting process more efficient. Computational models of visual attention attempt to mimic the behavioral aspects of the HVS. The proposed models in the literature may belong to three different branches namely bottom-up, top-down, or hybrid models combining the previously.

(a) Bottom-up

Bottom-up mechanisms are agnostic to the task at hand and have the purpose of extracting relevant low-level features and finding the most salient regions where attention should be deployed.

The pioneering works of Koch and Ullman (1987), Itti et al. (1998) combine multi-scale low-level features into a single saliency map. At first, spatial feature maps are built by extracting prominent local features from different feature modalities (color, intensity, orientation), using center-surround operations at different scales. Then, each map is normalized and linearly combined in a single saliency map. Finally, the Winner Take All (WTA) principle is applied to select the most salient locations to be sequentially analyzed, in order of decreasing conspicuity, using an Inhibition of Return (IOR) mechanism (Tipper et al., 1991).

Osberger's approach (Osberger & Maeder, 1998) starts by performing image segmentation and then assigning perceptual importance based on low-level image features—contrast, size, shape, color and motion—and high-level features—location, people and context. Osberger chose only 5 features to use in his algorithm and, per region, assigns an importance score to each. Lastly, a combination of these features results in a map which represents important regions in an image. Kadir and Brady (2001) identify salient regions based on entropy measures of image intensity while Gao and Vasconcelos (2007) defined a salient region considering how different this is from the surrounding background (center-surround mechanism (Siagian & Itti, 2007)).

The method of Gao et al. (2018) proposes a reinforcement learning framework for coarse-to-fine object detection. The method starts by applying an object detector at a down-sampled version of the original images, then on higher-resolution regions, that are likely to increase object detection accuracy. More specifically, the approach utilizes detection estimates to predict the accuracy gain for analyzing a region at a higher resolution (R-model) and a model that sequentially selects regions to zoom in (Q-net). The approach maintains high detection accuracy on the YFCC100M dataset while reducing the number of processed pixels by about 70% and the detection time by over 50%.

(b) Top-down

The top-down models take into account the observer's prior knowledge, expectations and current goals. The literature on visual attention suggests several sources of top-down influences (Borji & Itti, 2013b) when the problem is to decide which stimuli is important: attention can be drawn to specific object visual features in search models to easily reach the goal or use the context or gist to constrain search locations. Whenever there is a search task, top-down processes tend to dominate guidance and target-specific features are an essential source to draw attention more effectively. Moreover, our attention is oriented to task-relevant features. This way, attentional resources are not wasted and time and computational effort are saved for processing more pertinent/relevant parts of the visual field. Under these conditions, one knows what is looking for (goal) and we know from a priori knowledge to distinguish the features that we should be searching for. Thereby as defended by guided search theory (Wolfe et al., 1989; Wolfe, 1994), we are able to modulate the gains assigned to different features. If, for example, the task is to find a green object, the gain assigned to green color will be higher.

(c) Hybrid

Most visual attention approaches, model bottom-up and top-down processes independently. However, there must be a trade-off between purely bottom-up models that typically miss to detect inconspicuous objects of interest and top-down

systems that confine visual understanding according to prior expectations related to the task.

In recent years, a combination of bottom-up and top-down models, that we designate as hybrid models, have been presented. For instance, Frintrop's model (Frintrop, 2006) is compound by two saliency maps: one corresponding to top-down influences and another related with bottom-up influences. The aggregated saliency map is computed as a linear combination of those maps using a fixed weight which revealed to be a non-flexible approach. Rasolzadeh et al. (2007) presented a more flexible model where the combination of top-down and bottom-up saliency maps is done dynamically, using entropy measures that provide information of how the linear combination of weights should change over time. Conspicuity maps were created following Itti's approach in Itti et al. (1998) besides the extra parameters used to weight the saliency map. They used a neural network to learn the bias of the top-down saliency map based on information provided by contextual scene and the current task. These hybrid models suggest that the HVSs can guide attention by applying top-down weights on bottom-up saliency maps allowing quicker target detections in backgrounds full of distractors (Rasolzadeh et al., 2007). The authors in Zhang et al. (2008) proposed a probabilistic Bayesian framework for saliency learning using natural statistics (SUN). The most salient features are the ones with the highest point-wise self-information from features prior learned from a set of natural images, i.e., features that mostly differ from the learned average and are statistically unexpected (bottom-up modulation), or have the highest mutual information when searching for a specified target object (top-down modulation).

3 Attention in visual understanding tasks

Object classification consists of assigning a single label to a given image. Localisation includes not only classifying the subject of an image but also identifying its position, usually by means of a rectangular bounding box. Object detection assumes the possibility that more than a single instance can exist in a single image, namely of different classes. Thus the desired output consists of every instance's class label and respective bounding box. Classical methods for visual recognition tasks in the computer vision literature, extract key point features from the image, using hand-crafted filters, namely Histogram of Gradients (HOG) (Dalal & Triggs, 2005) or SIFT (Lowe, 1999). During a training phase, features are extracted from a set of different viewpoints, and stored in a database. In the online recognition phase, extracted features are matched against the database, based on their Euclidean distance. The implementation is typically a hash table and the Generalized Hough Transform (GHT) employed for fast and robust model matching. One successful example in the literature is the Aggregated Channel Features (ACF) of Dollár et al. (2012) for pedestrian detection, which employs a sliding window detection by classification approach, in which each window is binary classified as "person" or "not a person". Classification is performed using boosted decision trees, trained with labeled samples of full body pedestrians, using the Adaboost algorithm (Freund & Schapire, 1997). The classification method relies on handcrafted features that combine several image channels: LUV, Gradient Magnitude and HOGs channels aggregated in a blockwise manner. For multi-scale detection, the method uses multi-channel pyramids. The computational burden of constructing full pyramids is cleverly avoided by approximating in-between scales from interpolations of the coarser scales. Finally, non-maximum suppression is applied to avoid multiple detections (only a few pixels apart) that correspond to the same person (see Fig. 7a).

Recently, DNNs which are potent machine learning tools for pattern recognition inspired by neuronal network models in the brain, were developed to autonomously generate visual characteristic hierarchies. These can implicitly learn highly non-linear and non-convex functions, in an end-to-end manner, and hierarchical feature representations, optimized by training with large annotated datasets for recognizing complex patterns, circumventing the need of explicit feature engineering and selection. Deep learning techniques have been successful in different challenging visual tasks, not only on object detection (Redmon et al., 2016; Liu et al., 2016) (see Fig. 7b), but also on segmentation (He et al., 2017) and tracking (Held et al., 2016; Mnih et al., 2014), having recently surpassed humans in some classification tasks (He et al., 2015).

The aforementioned network architectures show the progress in object classification tasks. However, we have not yet addressed intuitively more challenging problems such as object detection.

Their proposed method entitled R-Convolutional Neural Network (CNN) (Long et al., 2015) first extracts region proposals from the image, and then feeds each region to a CNN with a similar architecture to that of AlexNet (Krizhevsky et al., 2012). The output of the CNN is then evaluated by a Support Vector Machine (SVM) classifier. Finally, the bounding boxes are tightened by resorting to a linear regression model. This network produces the set of bounding boxes surrounding the objects of interest and the respective classification. The region proposals are obtained through selective search (Uijlings et al., 2013). This method has a major pitfall—it is very slow. This is due to requiring the training of three different models simultaneously, namely the CNN to generate image features, the SVM classifier and the regression model to tighten the bounding boxes. Moreover, each region proposal requires a forward pass of the neural network.

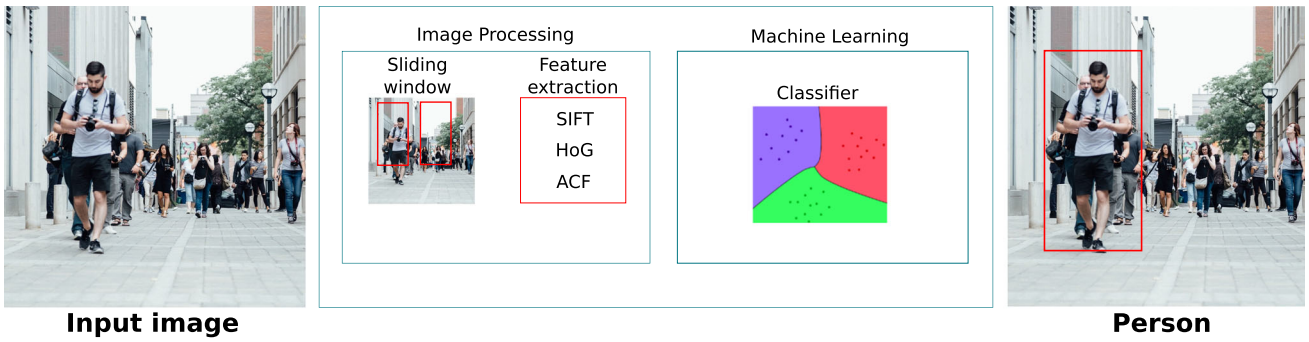


Fig. 7 Example architectures for visual object detection tasks. **a** Represents traditional methods in which hand-engineered features are fed to classical machine learning approaches. **b** Illustrates modern methods, based on simultaneously extracting features and regressing to bounding boxes, in an end-to-end manner, using deep learning based architectures

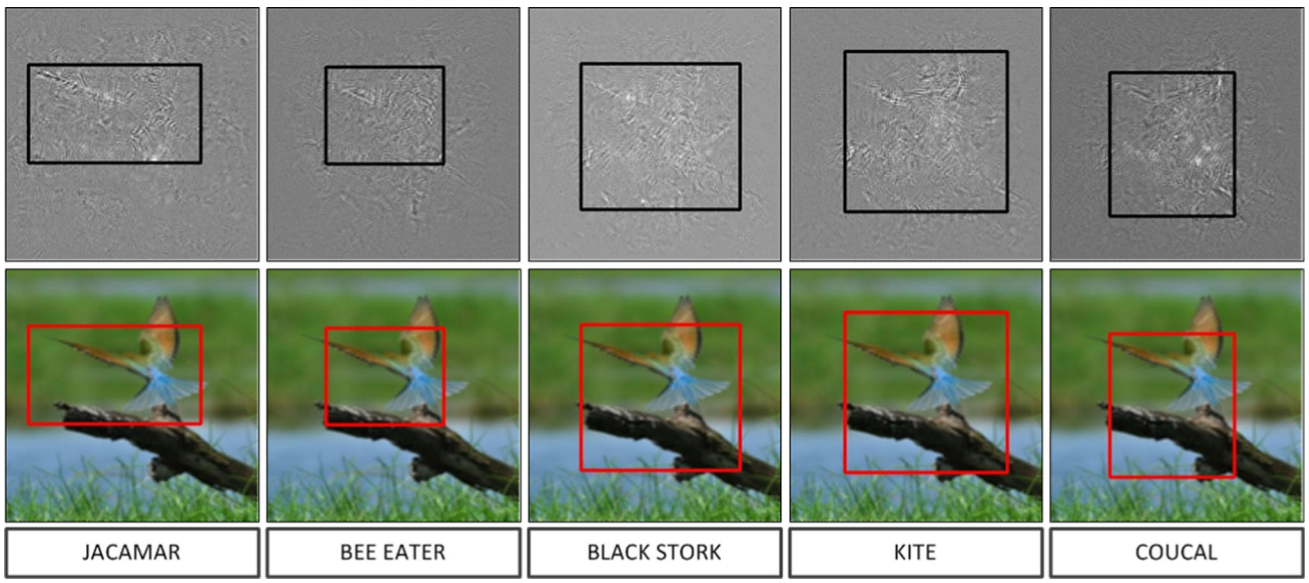


Fig. 8 Representation of the saliency map and the corresponding bounding box for each of the top-5 predicted class labels of a *bee eater* image of the ILSVRC 2012 data set (Russakovsky et al., 2015). The rectangles represent the bounding boxes that cover all non-zero saliency pixels resultant from a segmentation mask

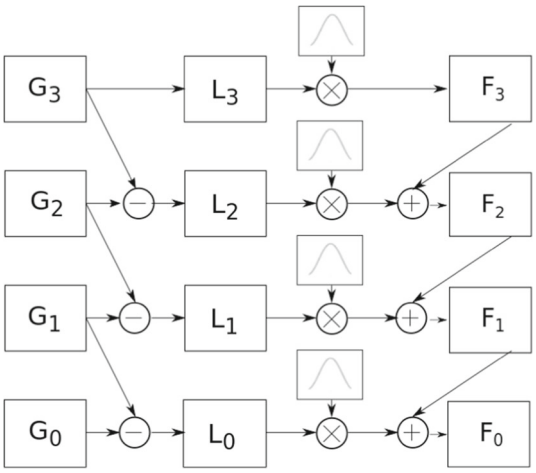


Fig. 9 A summary of the steps in the foveation system of Almeida et al. (2018) with four levels. The image G_0 corresponds to the original image and F_0 to the foveated image

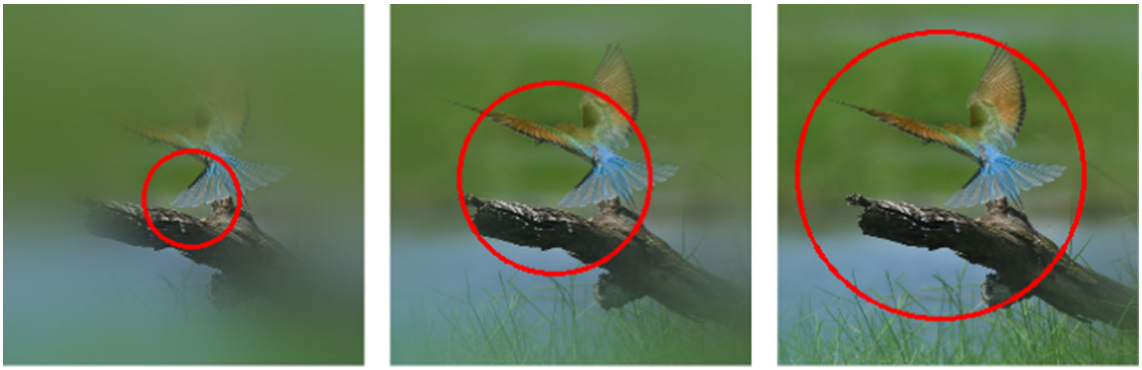
In 2015, Fast R-CNN (Girshick, 2015) was proposed to address the above-mentioned issues. This network has drastically faster performance and achieves higher detection quality. This is mainly due to two improvements: the first leverages the fact that there is generally an overlap between proposed interest regions, for a given input image. Thus, during the forward pass of the CNN it is possible to reduce the computational effort substantially by using Region of Interest (RoI) Pooling (RoIPool). The high-level idea is to have several regions of interest sharing a single forward pass of the network. Specifically, for each region proposal, we keep a section of the corresponding feature map and scale it to a pre-defined size, with a max pool operation. Essentially this allows us to obtain fixed-size feature maps for variable-size input rectangular sections. Thus, if an image section includes several region proposals we can execute the forward pass of the network using a single feature map, which dramatically speeds up training times. The second major improvement consists of integrating the three previously separated models into a single network. A Softmax layer replaces the SVM classifier altogether and the bounding box coordinates are calculated in parallel by a dedicated linear regression layer.

The progress of Fast R-CNN exposed the region proposal procedure as the bottleneck of the object detection pipeline. A Region Proposal Network (RPN) is a fully convolutional neural network (i.e. every layer is convolutional) (Ren et al., 2017) for simultaneously predicting objects' bounding boxes as well as objectness score. The latter term refers to a metric for evaluating the likelihood in the presence of an object of any class in a given image window. Since the calculation of region proposals depends on features of the image computed during the forward pass of the CNN, the authors merge RPN with Fast R-CNN into a single network, which was named Faster R-CNN. This further optimises runtime while achieving state of the art performance in the PASCAL VOC 2007, 2012 and Microsoft's COCO (Lin et al., 2014) datasets. However, the method is still too computationally intensive to be used in real-time applications, running at roughly 7 frames per second (FPS) in a high-end graphics card.

In the work of Almeida et al. (2018) the authors propose to capture visual attention through feedback Deep Convolutional Neural Network (DCNN). Their method uses a biologically inspired hybrid attention model, that combines bottom-up and top-down mechanisms and, additionally uses artificial human-like foveal vision, to efficiently locate and recognize objects in foveal digital images. More specifically, for a given input image I, the method computes a set of object class proposals by performing a feed-forward pass. The probability scores for each class label (N_c) are collected by accessing the network's output *softmax* layer. Then, retaining our attention on the five highest predicted class labels, then they compute the saliency map for each one of the predicted classes (see Fig. 8). Then, a top-down back-propagation pass is performed to compute the score derivative of the specific class c. The computed gradient indicates which pixels are more relevant for the class score (Simonyan et al., 2014). Figure 9 exemplifies the foveation model with four levels and Fig. 10 depicts examples of resulting foveated images. Kaplanyan et al. (2019) utilizes encoder-decoder networks that learn from sampled sparse video, a manifold of videos. It is trained on a large set of real-life videos, and uses recurrent convolutional neural networks that allows ensuring temporal stability of the reconstruction, by super-resolving features through time. The method is fast enough to be used in gaze-driven head-mounted real-time displays.

4 Resource-constrained perception in humanoid robotics

Space-variant vision and attention mechanisms have played a role of major importance in the design of energy and computational efficient robotics anthropomorphic heads (Rojas-Quintero & Rodríguez-Liñán, 2021). The Infanoid was an infant humanoid robot that featured efficient foveated stereo vision. The authors of Asfour et al. (2019) propose a humanoid robot for high performance complex tasks such as object manipulation, natural language understanding, integrated perception, and compliant motion-execution. It is equipped with a stereo camera system that has a baseline of $27cm$ and is used for foveal stereo active vision and a narrow one. The Karlsruhe humanoid head (Asfour et al., 2009) is a successful example in which foveal vision allows simple visuo-motor behaviors, such as smooth-pursuit and saccadic eye movements. Another example of mechanical head design is the work of Rojas-Quintero et al. (2021) that proposes a bio-inspired foveal and peripheral stereo vision



(a) $f_0 = 30$ (b) $f_0 = 60$ (c) $f_0 = 90$

Fig. 10 Example images obtained with the foveation system of Almeida et al. (2018) where $f_k = 2^k f_0$ defines the size of the region with highest acuity (the fovea), from a 227 × 227 uniform resolution image

system with kinematics that allow replicating saccadic movements. In the work of Adams et al. (2014), the authors propose a neuroanatomical model of visual attention in which objects in the surrounding environment of the cognitive agent (iCub (Sandini et al., 2007)) are attended depending on their event-driven (contrast change) bottom-up saliency (Galluppi et al., 2012), implemented using a Dynamic Video Sensor (DVS), and a neuromimetic biologically principled chip, SpiNNaker. The method of Ruesch et al. (2008) implements a multi-model saliency guided saccadic system in which not only visual (intensity, color, directional and motion features) but also sound cues are considered, and encoded in an ego-sphere to guide fixations.

5 Conclusions

In this article we have described the biological principles behind the human visual system and over-viewed approaches for biologically inspired artificial vision, ranging from low-level hardwired attention vision (i.e. foveal vision) to high-level visual attention mechanisms for robotics applications. More specifically, we over-viewed the state-of-the-art computational models for space-variant resource-constrained vision methods (foveal vision, selective attention, active vision), with application in important visual tasks (e.g. recognition and localization).

In particular, we have covered methods that show that biologically inspired selective attention mechanisms improve task execution, efficiency and speed, focusing on two important visual tasks: object recognition and localization. In the case of recognition, we emphasized approaches based on neural saliency mechanisms to actively center objects within the fovea through saccades. Finally, we over-viewed successful use-cases in robotics applications, namely anthropomorphic humanoid robotics heads, endowed with peripheral-foveal vision, active vision, and attention mechanisms.

Funding Open access funding provided by FCT|FCCN (b-on).

References

Adams, S. V., Rast, A.D., Patterson, C., Galluppi, F., Brohan, K., Pérez-Carrasco, J. A., Wennekers, T., Furber, S., & Cangelosi, A. (2014). Towards real-world neurorobotics: Integrated neuromorphic visual attention. In *International conference on neural information processing* (pp. 563–570). Springer.

Adelson, E. H., Anderson, C. H., Bergen, J. R., Burt, P. J., & Ogden, J. M. (1984). Pyramid methods in image processing. *RCA Engineer, 29*(6), 33–41.

Akbas, E., & Eckstein, M. P. (2017). Object detection through search with a foveated visual system. *PLoS Computational Biology, 13*(10), e1005743.

Almeida, A.F., Figueiredo, R., Bernardino, A., & Santos-Victor, J. (2018). Deep networks for human visual attention: a hybrid model using foveal vision. In A. Ollero, A. Sanfeliu, L. Montano, N. Lau, & C. Cardeira (Eds.), *ROBOT 2017: Third Iberian robotics conference* (pp. 117–128). Springer International Publishing. ISBN: 978-3-319-70836-2

Amso, D., & Scerif, G. (2015). The attentive brain: Insights from developmental cognitive neuroscience. *Nature Reviews Neuroscience, 16*(10), 606–619.

Asfour, T., Waechter, M., Kaul, L., Rader, S., Weiner, P., Ottenhaus, S., Grimm, R., Zhou, Y., Grotz, M., & Paus, F. (2019). ARMAR-6:

A high- performance humanoid for human-robot collaboration in real-world scenarios. *IEEE Robotics Automation Magazine, 26*(4), 108–121. https://doi.org/10.1109/MRA.2019.2941246

Asfour, T., Welke, K., Azad, P., Ude, A., Dillmann, R. (2008). The Karlsruhe humanoid head. In *Humanoids 2008—8th IEEE-RAS international conference on humanoid robots* (pp. 447– 453). https://doi.org/10.1109/ICHR.2008.4755993

Bailey, D. G., & Bouganis, C.-S. (2009). Vision sensor with an active digital fovea, 91–111

Bajcsy, R., Aloimonos, Y., & Tsotsos, J. K. (2018). Revisiting active perception. *Autonomous Robots, 42*(2), 177–196. https://doi.org/10.1007/s10514-017-9615-3

Balasuriya, S. L. (2006). A computational model of space-variant vision based on a self-organised artificial retina tessellation. Ph.D. thesis. University of Glasgow, UK. http://theses.gla.ac.uk/4934/

Balasuriya, S., & Siebert, P. (2005). A biologically inspired computational vision front-end based on a self-organised pseudorandomly tessellated artificial retina. In *Proceedings. 2005 IEEE international joint conference on neural networks* (Vol. 5, pp. 3069–3074). IEEE (2005)

Begum, M., & Karray, F. (2011). Visual attention for robotic cognition: A survey. *IEEE Transactions on Autonomous Mental Development, 3*(1), 92–105.

Bernardino, A., & Santos-Victor, J. (1999). Binocular tracking: Integrating perception and control. *IEEE Transactions on Robotics and Automation, 15*(6), 1080–1094.

Bisley, J. W. (2011). The neural basis of visual attention. *The Journal of Physiology, 589*(1), 49–57.

Bolduc, M., & Levine, M. D. (1998). A review of biologically motivated space-variant data reduction models for robotic vision. *Computer Vision and Image Understanding, 69*(2), 170–184.

Borji, A., & Itti, L. (2013b). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207. https://doi.org/10.1109/TPAMI.2012.89

Broadbent, D. (1958). *Perception and communication*. Pergamon Press.

Burt, P., & Adelson, E. (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications, 31*(4), 532–540.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research, 51*(13), 1484–1525. https://doi.org/10.1016/j.visres.2011.04.012

Cheung, B., Weiss, E., & Olshausen, B. A. (2017). Emergence of foveal image sampling from learning to attend in visual scenes. arXiv:1611.09430

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *IEEE computer society conference on computer vision and pattern recognition*, 2005. CVPR 2005 (Vol. 1, pp. 886–893). IEEE

de Figueiredo, R. P., Alexandre, B., Santos-Victor, J., & Araújo, H. (2018). On the advantages of foveal mechanisms for active stereo systems in visual search tasks. *Autonomous Robots, 42*(2), 459–476.

Dollár, P., Appel, R., & Kienzle, W. (2012). Crosstalk cascades for frame-rate pedestrian detection. In *Proceedings of the 12th European conference on computer vision—volume Part II. ECCV'12* (pp. 645–659). Springer. ISBN: 978-3-642-33708-6. https://doi.org/10.1007/978-3-642-33709-3_46.

Eriksen, C. W., James, St., & James, D. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics, 40*(4), 225–240.

Fernández-Caballero, A., & Ferrández, J.M. (2017). Biologically inspired vision systems in robotics.

Figueiredo, R., Dehban, A., Moreno, P., Bernardino, A., Santos-Victor, J., & Araújo, H. (2019). A robust and efficient framework for fast cylinder detection. *Robotics and Autonomous Systems, 117*, 17–28. https://doi.org/10.1016/j.robot.2019.04.002

Figueiredo, R., Moreno, P., & Bernardino, A. (2017). Robust cylinder detection and pose estimation using 3D point cloud information. In *2017 IEEE international conference on autonomous robot systems and competitions (ICARSC)* (pp. 234–239). IEEE.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–139.

Frintrop, S. (2006). *VOCUS: A visual attention system for object detection and goal-directed search*. Springer. ISBN 978-3-540-32760-8

Frintrop, S., Rome, E., & Christensen, H. I. (2010). Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception*. https://doi.org/10.1145/1658349.1658355

Galluppi, F., Brohan, K., Davidson, S., Serrano-Gotarredona, T., Pérez Carrasco, J. A., Linares-Barranco, B., & Furber, S. (2012). A real-time, event-driven neuromorphic system for goal-directed attentional selection. In *International conference on neural information processing* (pp. 226–233). Springer.

Gao, D., & Vasconcelos, N. (2007). Bottom-up saliency is a discriminant process. In *Proceedings of the IEEE international conference on computer vision*. https://doi.org/10.1109/ICCV.2007.4408851

Gao, M., Yu, R., Li, A., Morariu, V. I., & Davis, L. S. (2018). Dynamic zoom-in network for fast object detection in large images. In *2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 6926–6935). IEEE Computer Society. https://doi.org/10.1109/CVPR.2018.00724.

García, G., Jara, C., Pomares, J., Alabdo, A., Poggi, L., & Torres, F. (2014). A survey on FPGA-based sensor systems: Towards intelligent and reconfigurable low-power sensors for computer vision, control and signal processing. *Sensors, 14*(4), 6247–6278.

Geisler, W. S., & Perry, J. S. (1998). Realtime foveated multiresolution system for lowbandwidth video communication. In *Photonics West'98 electronic imaging* (pp. 294–305). International Society for Optics and Photonics.

Girshick, R. (2015). Fast R-CNN. In *2015 IEEE international conference on computer vision (ICCV)* (pp. 1440–1448). https://doi.org/10.1109/ICCV.2015.169

Gould, S., Arfvidsson, J., Kaehler, A., Sapp, B., Messner, M., Bradski, G., Baumstarck, P., Chung, S., & Ng, A.Y. (2007). Peripheralfoveal vision for real-time object recognition and tracking in video. In *Proceedings of the 20th international joint conference on artifical intelligence. IJCAI'07* (pp. 2115–2121). Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=1625275.1625617

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *2017 IEEE international conference on computer vision (ICCV)* (pp. 2980–2988). IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026–1034).

Healey, C. G., & Enns, J. T. (2011). Attention and visual perception in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics, 18*(7), 1–20.

Held, D., Thrun, S., & Savarese, S. (2016). Learning to track at 100 fps with deep regression networks. In *European conference on computer vision* (pp. 749–765). Springer.

Hinkle, D. A., & Connor, C. E. (2002). Three-dimensional orientation tuning in macaque area V4. *Nature Neuroscience, 5*(7), 665.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology, 148*(3), 574–591.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 160*(1), 106–154.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology, 195*(1), 215–243.

Itti, L., Koch, C., Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*(11), 1254–1259. https://doi.org/10.1109/34.730558. arXiv:0504378 [math]

Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. (2015). Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), Advances in neural information processing systems, vol. 28. Curran Associates, Inc.

James, W. (1890). The principles of psychology (Vols. 1 & 2). Holt 118 (p. 688). https://doi.org/10.1037/10538-000

Javier Traver, V., & Bernardino, A. (2010). A review of log-polar imaging for visual perception in robotics. *Robotics and Autonomous Systems, 58*(4), 378–398.

Kadir, T., & Brady, J. M. (2001). Scale, saliency and image description. *International Journal of Computer Vision, 45*(2), 83–105. https://doi.org/10.1023/A:1012460413855

Kaplanyan, A. S., Sochenov, A., Leimkühler, T., Okunev, M., Goodall, T., & Rufo, G. (2019). DeepFovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM Transactions on Graphics*. https://doi.org/10.1145/3355089.3356557

Kartheek Medathati, N. V., Neumann, H., Masson, G. S., & Kornprobst, P. (2016). Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision. *Computer Vision and Image Understanding, 150*, 1–30. https://doi.org/10.1016/j.cviu.2016.04.009

Katsuki, F., & Constantinidis, C. (2014). Bottom-Up and top-down attention: Different processes and overlapping neural systems. *The Neuroscientist, 20*(5), 509–521.

Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: towards the underlying neural circuitry. Matters of intelligence (pp. 115–141). Springer.

Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 25* (pp. 1097– 1105). Curran Associates, Inc.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Lawrence, C. (2014). ZitnickMicrosoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A. C. (2016). SSD: Single shot multibox detector. In *European conference on computer vision* (pp. 21–37). Springer.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *The proceedings of the seventh IEEE international conference on Computer vision* (Vol. 2, pp. 1150–1157). IEEE.

Lukanov, H., König, P., & Pipa, G. (2021). Biologically inspired deep learning model for efficient foveal-peripheral vision. *Frontiers in Computational Neuroscience*. https://doi.org/10.3389/fncom.2021.746204

Melício, C., Figueiredo, R., Almeida, A. F., Bernardino, A., & Santos-Victor, J. (2018). Object detection and localization with artificial foveal visual attention. In *2018 Joint IEEE 8th international conference on development and learning and epigenetic robotics (ICDL-EpiRob)* (pp. 101–106). https://doi.org/10.1109/DEVLRN.2018.8761032

Metta, G., Sandini, G., Vernon, D., Natale, L., & Nori, F. (2008). The iCub humanoid robot: An open platform for research in embodied cognition. In *Proceedings of the 8th workshop on performance metrics for intelligent systems* (pp. 50–56). ACM (2008)

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences, 6*, 414–417. https://doi.org/10.1016/0166-2236(83)90190-X

Mnih, V., Heess, N., & Graves, A., et al. (2014). Recurrent models of visual attention. In *Advances in neural information processing systems* (pp. 2204–2212).

Mohlin, C., Sandholm, K., Ekdahl, K. N., & Nilsson, B. (2017). The link between morphology and complement in ocular disease. *Molecular immunology, 89*, 84–99.

Osberger, W., & Maeder, A.J. (1998). Automatic identification of perceptually important regions in an image. In *Proceeding of the fourteenth international conference on pattern recognition* (Vol. 1, pp. 701–704). https://doi.org/10.1109/ICPR.1998.711240.

Parasuraman, R., & Yantis, S. (1998). *The attentive brain*. MIT Press.

Posch, C. (2012). Bio-inspired vision. *Journal of Instrumentation, 7*(01), C01054.

Posner, M.I. (2012). Cognitive neuroscience of attention. Guilford Press. ISBN: 9781609189853. http://books.google.pt/books?id=8yjEjoS7EQsC

Posner, M. I. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32*(1), 3–25. https://doi.org/10.1080/00335558008248231

Potapova, E., Zillich, M., & Vincze, M. (2017). Survey of recent advances in 3D visual attention for robotics. *The International Journal of Robotics Research, 36*(11), 1159–1176. https://doi.org/10.1177/0278364917726587

Quinlan, P., & Dyson, B. (2008). Attention: General introduction, basic models and data. *Cognitive Psychology*. https://doi.org/10.1136/ewjm.172.2.83

Rasolzadeh, B., Targhi, A.T., & Eklundh, J.-O. (2007). An attentional system combining top-down and bottom-up influences. In *Attention in cognitive systems. Theories and systems from an interdisciplinary viewpoint lecture notes in computer science* (Vol. 4840, pp. 123–140). https://doi.org/10.1007/978-3-540-77343-6_8. http://www.springerlink.com/index/682P7080741754X3.pdf

Recasens, A., Kellnhofer, P., Stent, S., Matusik, W., & Torralba, A. (2018). Learning to zoom: a saliency-based sampling layer for neural networks. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 51– 66).

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Rojas-Quintero, J. A., & Rodríguez-Liñán, M. C. (2011). A literature review of sensor heads for humanoid robots. *Robotics and Autonomous Systems, 143*, 103834. https://doi.org/10.1016/j.robot.2021.103834

Rojas-Quintero, J. A., Rojas-Estrada, J. A., Rodréguez-Sánchez, E. A., Vizcarra-Corral, J. A. (2021). Designing a bio-inspired foveated active vision system. In *2021 XXIII Robotics Mexican Congress (ComRob)* (pp. 1–6). https://doi.org/10.1109/ComRob53312.2021.9628636

Roncone, A., Pattacini, U., Metta, G., & Natale, L. (2016). A Cartesian 6-DoF Gaze controller for humanoid robots. Robotics: science and systems (Vol. 2016).

Rosenberg, A., Cowan, N. J., & Angelaki, D. E. (2013). The visual representation of 3D object orientation in parietal cortex. *Journal of Neuroscience, 33*(49), 19352–19361.

Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J., & Pfeifer, R. (2008). Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub. *IEEE International Conference on Robotics and Automation, 2008*, 962–967. https://doi.org/10.1109/ROBOT.2008.4543329

Russakovsky, O., Deng, J., Hao, S., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV), 115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Sandini, G., Metta, G., & Vernon, D. (2007). The iCub cognitive humanoid robot: An open-system research platform for enactive cognition. 50 years of artificial intelligence (pp. 358–369). Springer.

Santos-Victor, J., & Bernardino, A. (2003). Vision-based navigation, environmental representations and imaging geometries. Robotics Research (pp. 347–360). Springer.

Schwartz, E. L. (1977). Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception. *Biological Cybernetics, 25*(4), 181–194.

Schwartz, E. L., Greve, D. N., & Bonmassar, G. (1995). Space-variant active vision: Definition, overview and examples. *Neural Networks, 8*(7), 1297–1308. https://doi.org/10.1016/0893-6080(95)00092-5

Siagian, C., & Itti, L. (2007). Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(2), 300–312. https://doi.org/10.1109/TPAMI.2007.40

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. Computer Vision and Pattern Recognition. arXiv:1312.6034.

Sokolov, E.N., & Vinogradova, O.S. (1975). Neuronal mechanisms of the orienting reflex. L. Erlbaum Associates. ISBN: 9780470925621. https://books.google.pt/books?id=T1Z9AAAAIAAJ

Thavamani, C., Li, M., Cebron, N., & Ramanan, D. (2021). FOVEA: foveated image magnification for autonomous navigation. In *IEEE/CVF international conference on computer vision (ICCV)* (Vol. 2021, pp. 15519–15528). https://doi.org/10.1109/ICCV48922.2021.01525

Tipper, S. P., Driver, J., & Weaver, B. (1991). Object-centred inhibition of return of visual attention. *The Quarterly Journal of Experimental Psychology, 43*(2), 289–298.

Treisman, A. M. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*, 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Treisman, A. (1985). Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing, 31*(2), 156–177. https://doi.org/10.1016/S0734-189X(85)80004-9

Tsotsos, J. K. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Sciences, 13*(3), 423–445.

Tsutsui, K.-I., Taira, M., & Sakata, H. (2005). Neural mechanisms of threedimensional vision. *Neuroscience Research, 51*(3), 221–229.

Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., & Smeulders, A. W. M. (2013). Selective Search for Object Recognition. *International Journal of Computer Vision, 104*(2), 154–171. https://doi.org/10.1007/s11263-013-0620-5

Uzkent, B., & Ermon, S. (2020). Learning when and where to zoom with deep reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12345–12354).

Vijayakumar, S., Conradt, J., Shibata, T., & Schaal, S. (2001). Overt visual attention for a humanoid robot. In *2001 IEEE/RSJ international conference on intelligent robots and systems, 2001. Proceedings* (Vol. 4, pp. 2332–2337). IEEE.

Von Helmholtz, H. (1866). Handbuch der physiologischen Optik (Vol. 9).

Wang, Z. (2003). Rate scalable foveated image and video communications, Ph.D. thesis.

Wolfe, J. M. (1994). Guided Search 2.0 A revised model of visual search. *Psychnomic Bulletin & Review, 1*(2), 202–238. https://doi.org/10.3758/BF03200774

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 419–433.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision, 8*(7), 32–32.

Rui Pimentel de Figueiredo I have a Ph.D. degree in Electrical and Computer Engineering from Instituto Superior Técnico, VisLab/ISR. My research work at has been mostly related with 3D geometry processing, computer vision and robotics, with a strong emphasis on the subject of object recognition and pose estimation, using non-conventional human-like foveal vision. I was responsible for research and development, within two EU projects (First-MM and Handle) which were mainly directed towards vision for robot grasping and in-hand manipulation applications. In the context of these two projects I researched and developed novel state-of-the-art algorithms for object detection, recognition and pose estimation, that were deployed in robotic platforms operating in real environments. In the meantime, I have strong object oriented programming languages skills (C++ and Python). I have also become very familiar with the OpenCV and the Point Cloud Library (PCL) for perception, and the Robot Operating System (ROS) motion planning (MoveIt!) and navigation stacks. I have recently obtained a PhD in Electrical and Computer Engineering, with an emphasis in computer vision and robotics, under the Robotics, Brain and Cognition (RbCog) FCT program. The areas covered include: sensorimotor coordination, multisensory fusion, attention, human gesture recognition, uncertainty modelling, and learning from demonstration, with an emphasis on studying and developing novel biologically inspired attention mechanisms for resource-constrained perception. During this period I have published more than 15 works in top international journals and conferences, and co-supervised 4 MSc thesis. I am now a researcher at AirLab, Aarhus University, working on the design of intelligent computer vision based solutions for autonomous drone applications, with an emphasis on simultaneous localization and mapping (SLAM) for unmanned inspection tasks.

Alexandre Bernardino is an Associate Professor at the Dept. of Electrical and Computer Engineering of IST-Lisboa and Senior Researcher at the Computer and Robot Vision Laboratory of the Institute for Systems and Robotics of IST-Lisboa. He has participated in several national and international research projects as principal investigator and technical manager. He published more than one hundred research papers on top journals and peer-reviewed conferences in the field of robotics, vision and cognitive systems. He is associate editor of the journal Frontiers in Robotics and AI and of major robotics conferences. He is the chair or the IEEE Portugal RAS Chapter. His main research interests focus on the application of computer vision, machine learning, cognitive science and control theory to advanced robotics and automation systems.